Title	遺伝的アルゴリズムにおける平均最短距離の導出 (生命現象と関連した非線形問題の数理)
Author(s)	船谷, 浩之; 池田, 和司
Citation	数理解析研究所講究録 (2008), 1616: 137-144
Issue Date	2008-10
URL	http://hdl.handle.net/2433/140148
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
第1616巻 2008年 137-144

遺伝的アルゴリズムにおける平均最短距離の導出

船谷 浩之，池田和司
京都大学情報学研究科

概要: 近年，スモールワールド性と，探索の効率性の関連が指摘されている．本稿では，GAの交叉が探索の効率に及ぼす影響を評価するために，世代同士の距離を定義し，スモールワールド性の指標である平均最短距離を計算する．特に交叉に注目し，突然変異のみのGAと，交叉も入れたGAで比較を行う．遺伝子長 L，遺伝子数 $n=2$ の場合にそれぞれの平均最短距離の厳密解を，$n \geq 3$ の場合に近似解を導出した．

キーワード: 遺伝的アルゴリズム, スモールワールド, 平均最短距離，交叉

1 はじめに

GA(遺伝的アルゴリズム) は進化の過程を模倣した準最適化アルゴリズムである．GAの解析については，Hollandのスキーマ定理 [1,2] によって強い部分を持った集団が確率的に山を登る事が示され，マルコフ連鎖による解析 [3,4] によって収束が示されたが，収束の速さに関しての議論は少ない．特に交叉については，アルゴリズムにおいて重要な役割を果たすとされてきたにもかかわらず，その影響について定量的な解析が無いのが現状である．

本稿では，交叉の影響を評価するために，ネットワーク理論を用いる．スモールワールド性 [5] は現実世界のネットワークに多く見られ，リンク数が比較的少なく，ノード間の平均最短距離が小さい，という二つの性質を同時に満たす．平均最短距離 (Characteristic Path Length, CPL) とは，すべてのノードのペアについて，ノード間の最短距離の平均を取ったものである．近年，この性質を持ったネットワークの探索性について研究が行われている [6]．スモールワールド性と探索の効率性の関連は現在明らかでないが，関連性があると考えられる．特に交叉は，ネットワーク的に見れば，突然変異のみの規則的なネットワークに新たに枝を加えるオペレータであるので，ネットワーク的な解析により定量的な評価が出来るであろう．

以上のような動機から，GAがスモールワールド性を持っているかどうか，集団同士の距離を定義し，平均最短距離を調べる事により評価する事にする．まず $n=2$ の時を考え，次に n が一般の時についても評価する．

2 GAと遷移ネットワーク

ここでは，GAとその遷移ネットワークを導入する．ある問題の解の候補を個体，固体をコード化したものを遺伝子と呼び．遺伝子は通常 0 と 1 のビット列で表現される．遺伝子の集団を遺伝子群と呼び，個々を遺伝子，遺伝子の中の特定の位置を遺伝子座と呼ぶ．アルゴリズムは次のように実行される (図 1)．

1. 初期集団の生成
2. 適合度の計算
3. 遺伝子の選択 (淘汰)
4. 交叉，突然変異により次の世代を生成する
5. 終了条件が満たしていれば終了，そうでなければ，2-4 を繰り返す．

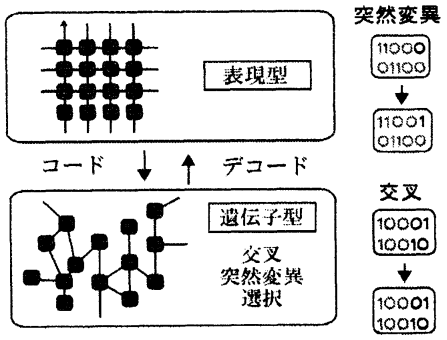

図 1: GA

本稿ではバイナリコーディングを用いる．まず，遺伝子数 $n=2$，遺伝子の長さ L，交叉，突然変異のみのGAを考える．突然変異は $2L$ ビットの中から，交叉は $L-1$ 個の候補の中から行われるものとする．この場

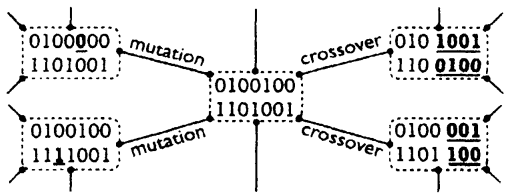

図 2: GA ネットワークの一部

合図 2 のように, 2 つの遺伝子を持つ個体群が, 突然変異もしくは交叉によってお互いに変化しながら, 世代交代を繰り返していく. 今, このネットワークを $\mathbf{G}_L$ とおく. $\mathbf{G}_L$ 上の 2 ノード間の距離は, 片方からもう一方へ $\mathbf{G}_L$ 上でリンクを辿っていった時の重みの合計として定義される. 異なる個体群, すなわちノードの数は

$$N \equiv 2^{L-1}(2^L + 1) \tag{1}$$

個あり, 異なる個体群のペアの数は全部で

$$M \equiv 2^{N-1}(2^N + 1) \tag{2}$$

個ある. また, すべてのノードから $k_m \equiv 2L$ 個の突然変異によるリンクがあり, 個体群を変化させる交叉が, 1 つのノードに k_c 個あるとしてその平均は,

$$\mathbb{E}[k_c] \equiv \frac{1}{4}\left(L - 1 + \frac{1}{2^{L-1}}\right) \tag{3}$$

となる (付録 1). ただし, $\mathbb{E}[\cdot]$ はここでは, ノードに関する平均を示す. 交叉による変化と, 突然変異による変化は必ず違う個体群を作るので, $\mathbf{G}_L$ のリンクの数は,

$$K \equiv k_m + \mathbb{E}[k_c] = N\left(\frac{9}{4}L - \frac{1}{4} + \frac{1}{2^{L+1}}\right) \tag{4}$$

となる.

3 平均最短距離 ($n = 2$ の場合)

3.1 交叉と突然変異の性質

ネットワーク $\mathbf{G}_L$ のスモールワールド性を調べるために, CPL を導出する. 以後, 常に 2 つの個体群に対してその最短距離を考察するので, 図 3 のようにそれぞれを P_o, P_d $(1 \le o, d \le N)$ と呼び, それぞれに属する合計 4 つの遺伝子を, $g_{o1}, g_{o2}, g_{d1}, g_{d2}$ と呼ぶ. 遺伝子のビット位置を遺伝子座, ある遺伝子座を $g_{o1}, g_{o2}, g_{d1}, g_{d2}$ の順に並べたものを遺伝子座パターンと呼ぶ. CPL を導出するにあたって, 遺伝子座パターンが重要になる.

図 4 に遺伝子座パターンの分類を示す. T_1, T_2, T_3 はそれぞれ最短 0, 1, 2 回の突然変異によって P_o, P_d が一致する. これらを突然変異パターンと呼ぶ. また, T_4, T_4' は他の遺伝子座パターンに影響され, 一回の交叉で 2 つ以上のビットを一致させる事が出来る. これらを交叉パターンと呼ぶ. 以降, T_1, T_2, T_3, T_4, T_4' を用いて, 個体群ペア $\mathbf{T}_{od}$ を表現する. 例えば図 3 に現れる個体群ペアは, $\mathbf{T}_{od} = \{T_2T_4'T_4T_2T_4T_3T_4\}$ である. 誤解の恐れが無ければ単に $\mathbf{T}$ と印す. ここで, 次の諸定理が成り立つ.

定理 1 突然変異は, 他の遺伝子座パターンに影響しない.

証明: 突然変異の性質を見れば明らかである.

定理 2 交叉は, 他の突然変異パターンには影響しない.

証明: もしある交叉地点で交叉をしても, 他の突然変異パターンにおいて P_o が P_d に一致するための突然変異回数が変わらないという事である. T_1, T_3 は *2* 遺伝子のビットが同じなので, 交叉の影響を受けない. T_2 の一部に交叉によってパターンが変わるものがあるが, それは交叉後もやはり T_2 であり, 突然変異回数は変わらない.

定理 3 重みが同じならば, *1* 回の有効な交叉は, *1* 回の有効な突然変異と同じか, それ以上距離を縮める. ここで有効とは, 個体群ペアの距離が縮まる事を示す.

証明: 交叉が, 交叉地点から右の遺伝子をすべて入れ替える事を考えれば明らかである. *1* 回の有効な交叉と *1* 回の有効な突然変異が同じ場合とは, 次のような例を考えればよい. 部分パターン $\{T_4T_4'T_4\}$ を $\{T_4T_4T_4\}$ に一致させたい場合, T_4' の両側で *2* 回の交叉を行う場合と, T_4' を二回の突然変異で T_4 に変化させるには, 同じ距離 *2* が必要である.

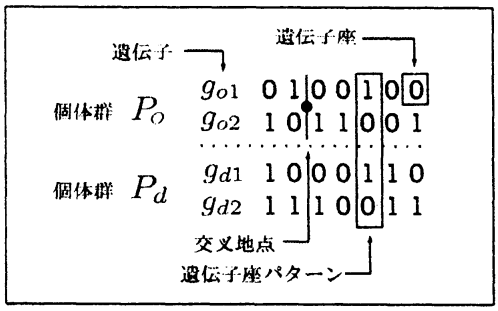

図 3: 個体群ペアと用語

3.2 最短距離 c_{ij} の導出

では, 任意の個体群ペア $\mathbf{T}_{od}$ $(o \neq d)$ の最短距離 c_{od} を求めよう. 定理 1–3 より, T_1, T_2, T_3 と, T_4, T_4' は別々に考える事が出来る. ここで, T_i が $\mathbf{T}$ において ℓ_i 回現れるとする. ℓ_4 は T_4, T_4' を合わせた長さとする. この時明らかに,

$$L = \sum_i \ell_i \tag{5}$$

であり, T_1, T_2, T_3 を一致させるために必要な突然変異回数を考え, c_{od} のうち, 突然変異によるもの $\hat{c}_{od}$ は,

$$\hat{c}_{od} = \ell_2 + 2\ell_3 \tag{6}$$

である. 次に, $\mathbf{T}$ のうち T_4, T_4' のみを抜きだしたパターンを, $\mathbf{T^4}$ とおくと, $\mathbf{T^4}$ の長さは ℓ_4 である. 今, 例えば $\mathbf{T^4} = \{T_4T_4T_4'T_4T_4'T_4'T_4\}$ の場合を考えると, 必要な最小交叉回数は 4 回である. これは, 左から見ていった時の T_4, T_4' が変化する回数を求めればよい. ビット列を扱う時に, ビット変化に排他的論理和を用いる事を思いだすと, $\mathbf{T^4}$ を一時的にビット $\{\hat{0}\hat{1}\}$ で表現し, 1 ビットシフトし, 排他的論理和を取ったもの $\mathbf{X}_{od}$ の 1 の数が, 求める最小交叉回数である (図 5). この数を $\tilde{c}_{od}$ として, 任意の $\mathbf{T}$ の最短距離は,

$$c_{od} = \hat{c}_{od} + \tilde{c}_{od} \tag{7}$$

となる.

3.3 平均最短距離の導出 (突然変異のみの場合)

CPL は, c_{od} の平均であり, 次のように定義される.

$$C \equiv \mathbb{E}[c_{od}] = \frac{1}{M} \sum_{o \neq d} c_{od} \tag{8}$$

$\mathbb{E}[\cdot]$ はこれ以後すべての $\mathbf{T}_{od}$ に対する平均を示す. 二つの遺伝子群の単純な距離, つまり 2 つの $2L$ ビット列

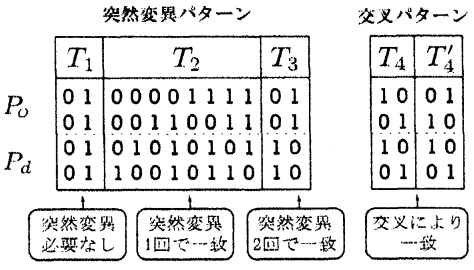

図 4: 遺伝子座パターン

$$\begin{array}{rl} & \boxed{T_4 \to \hat{0} \quad T_4' \to \hat{1}} \\ \hat{0}\hat{0}\hat{1}\hat{0}\hat{1}\hat{1}\hat{0} & \leftarrow \{T_4T_4T_4'T_4T_4'T_4'T_4\} \\ \text{xor}\ \ \hat{0}\hat{0}\hat{1}\hat{0}\hat{1}\hat{1}\hat{0} & \leftarrow \text{1 bit shift} \\ \hline \hat{0}\hat{1}\hat{1}\hat{1}\hat{0}\hat{1} & \leftarrow \mathbf{X}_{od} \end{array}$$

図 5: 最小交叉数の求め方

のハミング距離の平均は, L である事は明らかである. しかし今, 遺伝子を群として考えているので, 単純に二つの遺伝子の距離を取っただけでは最短距離は求まらない. これは, ペアの選び方が複数あり, その距離の合計を最小にする必要があるためである. ここで, 16 個の遺伝子座パターンに注目する. 突然変異のみの遷移を考える場合, 遺伝子のペアの取り方に関係するのは, T_4, T_4' である. 他のパターンは, ペアの取り方によらず必要な突然変異回数は変わらない. 遺伝子群同士の突然変異を最小にするには, この二つのパターンの少ない方を突然変異してやればいい事になる. 図 6 では, T_4' の数のほうが少ないため, $g_{o1} \to g_{d1}, g_{o2} \to g_{d2}$ という風に一致させてやれば, 最小の突然変異回数で二つの遺伝子群が一致する事になる. 次に実際に必要な突然変異回数を計算しようえ.

二項分布を正規分布で近似するとして, 遺伝子長 L の遺伝子群ペアの中の T_4 の数 Y は,

$$Y \sim \mathcal{N}\left(\frac{L}{8}, \frac{7L}{64}\right)$$

に従う. $\mu_y \equiv \frac{L}{8}, \sigma_y^2 \equiv \frac{7L}{64}$ とおく. 求めたいものは, Y を 2 個 i.i.d で発生させたものの最小値 $\min\{Y_1, Y_2\}$ であるので, これを Z と置くと, Z の累積密度関数は,

$$\text{Prob}(Z \leq z) = 1 - \prod_{i=1}^{2} \text{Prob}(Z > z)$$

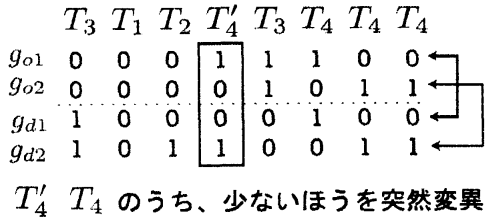

図 6: 突然変異のみの CPL の求め方

となるので，Z の確率密度関数は，

$$\begin{aligned}
&\mathrm{Prob}(Z=z)\\
=\ &\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{Prob}(Z\geq z)\\
=\ &\frac{\mathrm{d}}{\mathrm{d}z}\prod_{i=1}^{2}\int_z^\infty \frac{1}{\sqrt{2\pi\sigma_y^2}}e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}\mathrm{d}y\\
=\ &\frac{\mathrm{d}}{\mathrm{d}z}\left(\int_z^\infty \frac{1}{\sqrt{2\pi\sigma_y^2}}e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}\mathrm{d}y\right)^2\\
=\ &\frac{1}{\sqrt{2\pi\sigma_y^2}}e^{-\frac{(z-\mu_y)^2}{2\sigma_y^2}}\tilde{\Phi}\left(\frac{z-\mu_y}{\sqrt{2\sigma_y^2}}\right)
\end{aligned}$$

となる．ただし，$\tilde{\Phi}(z)$ は complementary error function

$$\tilde{\Phi}(z)=\frac{2}{\sqrt{\pi}}\int_z^\infty e^{-t^2}\mathrm{d}t$$

である．突然変異は Z の 2 倍必要であり，その他のパターンに必要な突然変異の平均は $\frac{3}{4}L$ なので，求める近似平均最短距離は，

$$C_m=\frac{3}{4}L+2\mathbb{E}[Z] \tag{9}$$

となる．

3.4 平均最短距離の導出 (交叉も含めた場合)

次に，交叉を含めた場合を考えよう．明らかに，突然変異の時のように T_4, T_4' を突然変異させるよりも，交叉したほうが距離は短くなる事が分かる．この時，T_4, T_4' 以外のタイプの分布は，組み合わせを考えない時と同じものと考えてよい．

さて，$\mathbb{E}[c_{od}]$ を求めるには，まず (5) を満たす ℓ_i の組みあわせをすべて考え，ℓ_4 に関して，図 5 に示した排他的論理和に注意して交叉回数を数えてやればよい [8]．以下では直接的に各パターンの割合を考える事により CPL を導出する．

突然変異については，定理 1 より遺伝子座パターンによって独立に考える事が出来る．今，$\mathbf{T}_{od}$ の中の一番左の遺伝子座パターンに着目する．すべての，$\mathbf{T}_{od}$ を考えた時，この部分の遺伝子座パターンには，図 4 における 16 種類のパターンが同じ割合で現れる．これは $\mathbf{T}_{od}$ のどの遺伝子座パターンを見ても同じである．すなわち，ある $\mathbf{T}_{od}$ が必要とする突然変異回数 $\hat{c}_{od}$ の平均は，

$$\mathbb{E}[\hat{c}_{od}]=\left(0+\frac{1}{2}+2\cdot\frac{1}{8}\right)L=\frac{3}{4}L \tag{10}$$

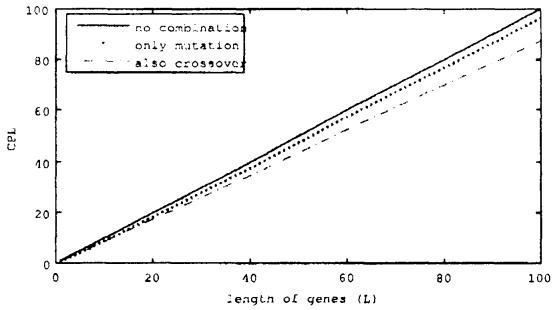

図 7: $n=2$ における平均最短距離

となる．

次に交叉について考える．図 5 の考察の結果，最小交叉回数 $\tilde{c}_{od}$ は排他的論理和 $\mathbf{X}_{od}$ により導出される．同様に割合を考えると，図 4 より T_4, T_4' は全体の $\frac{1}{4}$ であるので，ℓ_4 は一つの個体群あたり平均 $\frac{1}{4}$ になる．$\mathbf{X}_{od}$ は，$\ell_4=0$ の場合は定義されず，それ以外は ℓ_4-1 となる．つまり $\mathbf{X}_{od}$ の長さ $\mathbf{x}_{od}$ の平均は，

$$\mathbb{E}[\mathbf{x}_{od}]=\frac{L}{4}-\left\{1-\left(\frac{3}{4}\right)^L\right\} \tag{11}$$

である．$\mathbb{E}[\mathbf{x}_{od}]$ のうちの $\frac{1}{2}$ が 1 であり，交叉している事を示すので，結局 $\tilde{c}_{od}$ の平均は，

$$\mathbb{E}[\tilde{c}_{od}]=\frac{1}{2}\mathbb{E}[\mathbf{x}_{od}]=\frac{L}{8}-\frac{1}{2}\left\{1-\left(\frac{3}{4}\right)^L\right\} \tag{12}$$

となる．以上より，突然変異と交叉を含めた平均最短距離は，

$$\begin{aligned}
C_c\ &\equiv\ \mathbb{E}[\hat{c}_{od}]+\mathbb{E}[\tilde{c}_{od}] &(13)\\
&=\ \frac{7}{8}L-\frac{1}{2}\left\{1-\left(\frac{3}{4}\right)^L\right\} &(14)
\end{aligned}$$

となる．

図 7 に組み合わせを考えない時のハミング距離，突然変異のみの場合の CPL，交叉を含めた場合の CPL を示す．組み合わせを変えて突然変異するより，交叉を優先した方が CPL は短くなっている事が分かる．

4 確率的な重みの導入

ここまでの議論では，突然変異と交叉は同じ重みを持っていた．突然変異のみのネットワークに交叉が加わったのであるから，CPL が小さくなるのは当然である．そこで，突然変異確率 μ $(0<\mu\leq 1)$，交叉確率

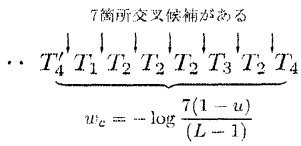

図 8: 確率を導入した時の交叉によるリンク重み

$1-\mu$ を定め，μ によって CPL がどう変化するのか見てみよう．

リンクの重みは，これが確率であり，ネットワーク上の距離が重みの足し算になっている事を考え，遷移確率 p の負対数を取って，

$$w \equiv -\log_2 p \tag{15}$$

と置くのが自然であろう．突然変異と交叉地点は，それぞれ等確率で選択されるとしているので，リンクの重みをそれぞれ w_m, w_c と置くと，

$$w_m(\mu) = -\log_2 \frac{\mu}{2L} \tag{16}$$

$$w_c(\mu, k) = -\log_2 \frac{k(1-\mu)}{L-1} \tag{17}$$

となる．k は，T_4, T_4' にの間ある T_1, T_2, T_3 の数である．すなわち，T_4, T_4' の間にあるそれ以外のパターンが多いほど，同じノードに到達する確率が高くなる (図 8)．

まず，世代交代が突然変異のみの場合の CPL を $C_m(\mu)$[1] とすると，

$$C_m(\mu) = -L \log \frac{1}{2L} = L \log 2L \tag{18}$$

となる．次に，交叉を含めたネットワークにおける CPL : $C_c(\mu)$ を考えよう．当然，定理 3 の「交叉と突然変異の重みが同じ」という仮定が成りたたなくなる μ が存在する．この時，最短距離パス上の交叉と突然変異が入れ替わってしまう．

では，μ がどんな値の時，突然変異が交叉に入れ替わるであろうか．それには，1 回の交叉と突然変異が同じ時を考えるればよい．この時，例えば $\mathbf{T}_{od}$ の部分パターンは $\{T_4 T_4' T_4\}$ になっているので，この時の μ を μ_k と置くと

$$\frac{\mu_k}{2L} = \frac{k(1-\mu_k)}{L-1} \quad \Leftrightarrow \quad \mu_k = \frac{2kL}{3kL+L-1} \simeq \frac{2k}{2k+1} \tag{19}$$

となる．ここで $L \gg 1$ としており，$u_1 \simeq 2/3$ である．$0 < \mu \leq \mu_1$ では，定理 3 が成りたつので，CPL は比較的簡単に計算可能である．次に 2 回の交叉と突然変異が同じ場合を考えよう．この時の μ は，

$$\left(\frac{\tilde{\mu}}{2L}\right)^2 = \frac{k(1-\tilde{\mu})}{L-1} \quad \Leftrightarrow \quad \tilde{\mu} \simeq 1 \tag{20}$$

となる．すなわち十分に長い L では，(19) の場合のみ考えてやればよい．

まず，常に交叉が選択される場合，すなわち $0 < \mu \leq \mu_1$ の場合の CPL : $C_{c1}(\mu)$ を考えよう．任意の $\mathbf{T}_{od}$ において突然変異による最短距離を $\hat{a}_{od}$ とし，交叉によるものを $\tilde{a}_{od}$ とする．$\hat{a}_{od}$ の平均は (10) の $\mathbb{E}[\hat{c}_{od}]$ と同じように考えられるので，

$$\mathbb{E}[\hat{a}_{od}] = -\frac{3}{4} L \log \frac{\mu}{2L} \tag{21}$$

である．交叉についても，重みの平均を考えてやればよい．今十分に長い L を考え，交叉にを起こす T_4, T_4' をランダムに選んだ時，その間の T_1, T_2, T_3 の数が k になる確率は $\frac{1}{4}\left(\frac{3}{4}\right)^k$ となるので，リンクの重み平均は，以下のように計算される．

$$\begin{aligned}
\mathbb{E}[w_c(\mu,k)] &= -\frac{1}{4}\sum_{k=0}^{L}\left(\frac{3}{4}\right)^k \log \frac{k(1-\mu)}{L-1} \\
&= -\frac{1}{4}\log\frac{(1-\mu)}{L-1}\sum_{k=0}^{L}\left(\frac{3}{4}\right)^k - \frac{1}{4}\sum_{k=0}^{L}\left(\frac{3}{4}\right)^k \log k \\
&\simeq -\log\frac{(1-\mu)}{L-1} - \gamma(L)
\end{aligned} \tag{22}$$

となる．ここで

$$\gamma(L) \equiv \frac{1}{4}\sum_{k=0}^{L}\left(\frac{3}{4}\right)^k \log k \tag{23}$$

であり，μ には依存しない数であり，$C(\mu)$ に対しては小さい値を取る．(7) において，重みを $\mathbb{E}[w_c(\mu,k)]$ に置きかえたものが交叉による最短距離となるので，

$$\mathbb{E}[\tilde{a}_{od}] \simeq \frac{L}{8}\left\{-\log\frac{(1-\mu)}{L-1} - \gamma(L)\right\} \tag{24}$$

となる．よって (21)，(24) により，$0 \leq \mu \leq \mu_1$ における CPL : $C_{c1}(\mu)$ は以下のようになる．

$$\begin{aligned}
C_{c1}(\mu) &\simeq \mathbb{E}[\hat{a}_{od}] + \mathbb{E}[\tilde{a}_{od}] \\
&= L\left(-\frac{3}{4}\log\frac{\mu}{2L} - \frac{1}{8}\log\frac{1-u}{L-1}\right) - \frac{L}{8}\gamma(L)
\end{aligned} \tag{25}$$

次に，$\mu > \mu_1$ の範囲について考える．μ を μ_1 から大きくしていくと，$w_c(\mu,k)$ における k が小さい交叉が突然変異に入れかわっていく．(25) を見ると，μ の変化により影響があるのは最後の項以外であるから，$k=1$

[1] μ は常に 1 なので，本来は C_m と書くべきだが，確率を考慮しない時の C_m と区別するためにこう書く．

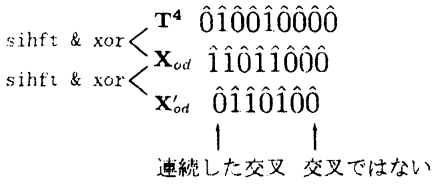

図 9: 連続して交叉が起こる場合

と近似し，$\mu = \mu_k$ で入れ替わる交叉がすべて，$\mu = \mu_1$ で入れ替わるとする．この時，$\mathbf{T_4}$ において，連続する交叉がどれくらいあるかを考えてやればよいが，このために $\mathbf{X}_{od}$ のさらに排他的論理和を取った $\mathbf{X}'_{od}$ を考える (図 9)．$\mathbf{X}'_{od}$ も $\mathbf{X}_{od}$ と同様，一様な $\mathbf{T}_{od}$ に対して $\hat{0}$ と $\hat{1}$ が均一に現れ，$\mathbf{X}'_{od}$ において $\hat{0}$ であり，そのすぐ上の $\mathbf{X}_{od}$ が 1 であるような時，連続した交叉となる．この時注意すべきは，連続する交叉のすべてが入れ替わるわけではなく，例えば $\{T_4T'_4T_4T'_4\}$ という風に T_4, T'_4 が交互に 3 回現れた場合，片方しか突然変異に替わらないという事である．つまり，「$\mathbf{T}$ の部分パターン $\mathbf{T}_4$ を考えた場合，連続して T_4, T'_4 が入れかわっている部分のうち，半分が入れ替わる」という事である．これは交叉のうち，$\frac{3}{4}$ が $\mu_1 \leq \mu$ で入れ替わる事を示す．以上を踏まえて，$\mu_1 \leq \mu \leq \tilde{\mu}$ における CPL : $C_{c2}(\mu)$ は，次のように近似される．

$$C_{c2}(\mu) \simeq -L\left(\frac{3}{4}+\frac{1}{8}\cdot\frac{3}{4}\right)\log\frac{\mu}{2L} - L\left(\frac{1}{8}\cdot\frac{1}{4}\right)\log\frac{1-u}{L-1} - \frac{L}{8}\gamma(L) \quad (26)$$

また，すべての交叉が突然変異に置き替わる時，$\mu = 1$ であり，$C_m(1) = C_c(1)$ となる事に注意されたい．

$$C_c(1) = L\log 2L \quad (27)$$

図 10 に $L = 50$ の時の μ vs. $C_c(\mu), C_m(\mu)$ のグラフを示す．$\mu = 1$ において，突然変異のみのネットワークと，交叉も含めたネットワークは一致し，$C(\mu)$ は $\mu \simeq \frac{27}{28}$ で極小値を取るような下凸関数となる．また $\hat{\mu}$ は $C_{c1}(\mu) = C_m$ となるような μ であり，これを μ について解くと，$\hat{\mu} \simeq 0.41$ となる．すなわち $C_c(\mu)$ と $C_m(\mu)$ は，次のような関係になる．

$$\begin{cases} C_c(\mu) \geq C_m(\mu) & (0 < \mu \leq \hat{\mu}) \\ C_m(\mu) \geq C_c(\mu) & (\hat{\mu} \leq \mu \leq 1) \end{cases} \quad (28)$$

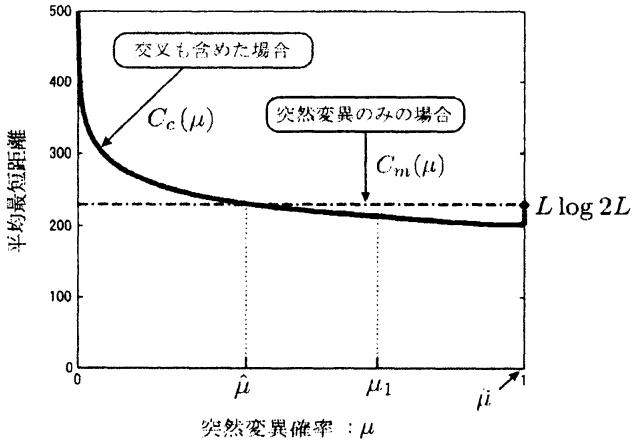

図 10: μ vs. $C(\mu)$; $L = 50$

5 平均最短距離 (n が一般の場合)

次に，遺伝子数 n が一般の場合を考えよう．$n = 2$ の時と同様に，突然変異のみの場合と交叉が入る場合を比べ，両者を比較する．

5.1 突然変異のみの場合

$n = 2$ の場合と同様に，n が一般の場合も n 個の遺伝子が群として同じならば，同じものとみなす．$n = 2$ の時は，遺伝子座パターンを見る事によりうまく計算出来たが，$n = 2$ の場合と違い，最適なペアリングを求める事が容易でないため，直接評価する事は難しい．

そこで，ある遺伝子 $\mathbf{g}$ と それ以外の $n-1$ 個の遺伝子のうち一番近いものの距離で近似する事にする．二項分布を正規分布で近似するとして，ランダムに発生させた二つの遺伝子のハミング距離距離 Y は，

$$Y \sim \mathcal{N}\left(\frac{L}{2}, \frac{L}{4}\right)$$

に従う．今求めたいものは，Y を n 個 i.i.d で発生させたものの最小値 $\min\{Y_i\}$, $i = 1, \cdots, n$ であるので，

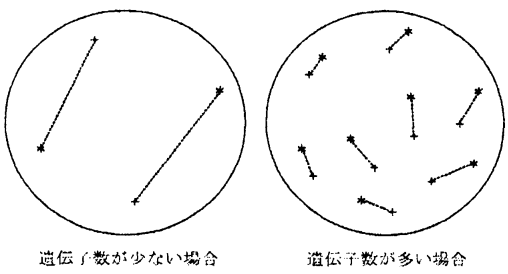

図 11: 遺伝子数が増えた場合の最短距離

これを Z と置くと，

$$\mathrm{Prob}(Z \leq z) = 1 - \prod_{i=1}^{n-1} \mathrm{Prob}(Y_i > z)$$

となるので，Z の確率密度関数は，$n = 2$ の時と同様に

$$\begin{aligned}
& \mathrm{Prob}(Z = z) \\
= & \frac{\mathrm{d}}{\mathrm{d}z}\mathrm{Prob}(Z \geq z) \\
= & \frac{\mathrm{d}}{\mathrm{d}z}\prod_{i}^{n}\int_{z}^{\infty}\sqrt{\frac{2}{L\pi}}e^{-\frac{2(y-L/2)^2}{L}}\mathrm{d}y \\
= & \frac{\mathrm{d}}{\mathrm{d}z}\left(\int_{z}^{\infty}\sqrt{\frac{2}{L\pi}}e^{-\frac{2(y-L/2)^2}{L}}\mathrm{d}y\right)^{n} \\
= & \frac{n}{\sqrt{L\pi}}e^{-\frac{2(z-L/2)^2}{L}}\left(\frac{1}{2}\tilde{\Phi}\left(\sqrt{\frac{2}{L}}\left(z-\frac{L}{2}\right)\right)\right)^{(n-1)}
\end{aligned}$$

となる．$n\mathbb{E}[Z]$ が求める近似平均最短距離となる．

5.2 交叉を含む場合の CPL 上界

次に交叉を入れた場合を考えよう．ここで我々は，2つの問題に直面するが，確率を含めた議論において，実際の GA は交叉の重みが突然変異に比べほとんど無い事を踏まえ，突然変異確率が小さい時の平均最短距離の上界を計算する．

問題の一つは，$N = 2$ の時と同じように，突然変異を選択するより交叉を選択する方が距離を短く出来るか，というもの．似たような系列があれば，それらをペアリングしてやれば，交叉の必要が無くなる．しかし，$N = 2$ の時に見たように，集団としての一致を考えた時の距離は，交叉に比べてあまり小さくならない．つまり，N が増加した時も，交叉を優先したほうが距離を縮めると考えられる．このように考えると，遺伝子の長さが 1 だと考えた時の必要な突然変異回数を L 倍する事によって，CPL が見積もれる事が分かる．

二つめは，計算機による実験がやりにくいという事である．交叉ポイントの選択は，一つの遺伝子集団に対し $O(2^{LN})$ 個の候補があり，突然変異は $O(LN)$ 個選択肢があるので，$O(LN2^{AN})$ の中から最短になるようなものを選択しなければならない．これらの平均をとるのは，計算機的にも難しい．

計算機実験の困難さはあるものの，上記のように考える事により，CPL の上界を求める事が出来る．つまり，$L = 1$ の時に必要な平均突然変異回数を求める．これは，平均 $n/2$，分散 $n/8$ のベルヌーイ分布に従う独立で同一な確率変数 X_1, X_2 の差の絶対値をとる確率変数

$$Y_1 = |X_1 - X_2| \tag{29}$$

を考え，この期待値を計算してやればよい．確率変数 X_1, X_2 を同じ平均と分散を持つ正規分布で近似すると，$X_1 - X_2$ は平均 0，分散 $n/4$ に従う正規分布となるので，

$$\begin{aligned}
\mathbb{E}[Y] &= 2\int_{0}^{\infty}\sqrt{\frac{n}{2\pi}}x\exp\left(-\frac{2x^2}{n}\right) \\
&= \sqrt{\frac{n}{2\pi}}
\end{aligned} \tag{30}$$

となり，$L\sqrt{\frac{n}{2\pi}}$ が求める期待値となる．すなわち，交叉が突然変異に対し確率が小さい場合には，平均最短距離は遺伝子数 n に対して，$O(\sqrt{\pi})$ の速さでしか大きくならない事が分かった．

6 結論と今後

遺伝子長 $L \in \mathbb{N}$，遺伝子数 $n \in \mathbb{N}$ の GA において，突然変異のみの GA と，交叉を含む GA の CPL を近似的に導出した．交叉を含む場合では，突然変異が交叉に比べ確率が小さい時に，近似的な距離の上界を導出した．この場合，CLP は遺伝子数 n に対して $O(\sqrt{n})$ の速さで大きくなる事が分かった．

この考察から言える事は，交叉オペレータを導入し，突然変異確率が小さい場面では，遺伝子の数 N を増やしても CPL はそれほど大きくならないという事である．もちろん，突然変異確率を増やせば CPL は絶対的に小さくなる (図 10)．しかし，GA は，遺伝子数を増加させた時でも CPL があまり大きくならないような場面でよく動く，という結果は，CPL を最適化アルゴリズムの評価指標として用いる場合の一つの基準になると考えられる．

今後は，モデル問題に対して CPL と平均的によい結果を出すパラメータと遺伝子数の関係を実験的に考察する．

参考文献

[1] J. Holland, *Adaptation in Natural and Artificial Systems*, MIT Press Cambridge, MA, USA, 1992.

[2] D. Goldberg *et al.*, *Genetic Algorithm in Search, Optimization and Machine Learning*, Reading, 1989.

[3] J. Suzuki, "A Markov Chain Analysis on Simple Genetic Algorithms," *IEEE Trans. on Systems, Man and Cybernetics*, vol.25, no.4, pp.655–659, 1995.

[4] Goldberg, D.E. and Segrest, P., "Finite Markov chain analysis of genetic algorithms," *Proceedings of the Second International Conference on Genetic Algorithms*, pp.1–8, 1987

[5] D. Watts and S. Strogatz, "Collective Dynamics of Small-world Networks.," *Nature*, vol.393, no.6684, pp.409–10, 1998.

[6] Rosvall, M. and Gronlund, A. and Minnhagen, P. and Sneppen, K., "Searchability of networks," it Phys. Rev E, vol.72, no.4, p.46117, 2007

[7] V. Latora and M. Marchiori, "Efficient Behavior of Small-world Networks," *Phys. Rev. Lett.*, vol.87, no.19, p.198701, Oct 2001.

[8] H. Funaya and K. Ikeda, "A Network Analysis of Genetic Algorithms," *IEICE Trans. on Information and Systems*, vol.90, no.6, pp.1002–1005, 2007.

[9] P. Stadler and C. Stephens, "Landscapes and Effective Fitness," *Theoretical Biology*, vol.8, no.4, pp.389–431, 2003.

[10] D. Wolpert, W. Macready, I. Center, and C. San Jose, "No Free Lunch Theorems for Optimization," *IEEE Trans. on Evolutionary Computation*, vol.1, no.1, pp.67–82, 1997.